ReTaT: A Unified Benchmark for Relation Extraction across Text and Table

Mohamed Ettaleb[1], Thibault Ehrhart[2], Nathalie Aussenac-Gilles[1], Yoan Chabot[3]
Mouna Kamel[1,4], Véronique Moriceau[1], Raphaël Troncy[2], Fanfu Wei[2]
(1) Univ Toulouse, Toulouse INP, CNRS, IRIT, Toulouse, France
(2) EURECOM, France
(3) Orange Research, France
(4) Espace-Dev, Université de Perpignan, France
yoan.chabot@orange.com, raphael.troncy@eurecom.fr, {firstname.lastname@irit.fr}

Abstract

While prior work in Information Extraction (IE) has focused on extracting information from either textual content or tables in isolation, they miss critical information that emerges only from their interplay. Indeed, tables may summarize facts that are sparse in the text, while text can disambiguate or elaborate on table entries. This complementarity may take the form of relations which are expressed across text and tables. In this context, we are interested in extracting such relations whose expression spans the two modalities. We propose this original task for which no reference evaluation corpora exists. Thus, we created ReTaT, a dataset that can be used to train and evaluate systems for extracting such relations. This dataset is composed of (table, surrounding text) pairs extracted from Wikipedia pages and has been manually annotated with relation triples. ReTaT is organized in three subsets with distinct characteristics: domain (business, telecommunication and female celebrities), size (from 50 to 255 pairs), language (English vs French), type of relations (data vs object properties), close vs open list of relation, size of the surrounding text (paragraph vs full page). We then assessed its quality and suitability for the joint table-text relation extraction task using Large Language Models (LLMs), at a time when LLMs have demonstrated their ability to extract relations from either text or tables in isolation.

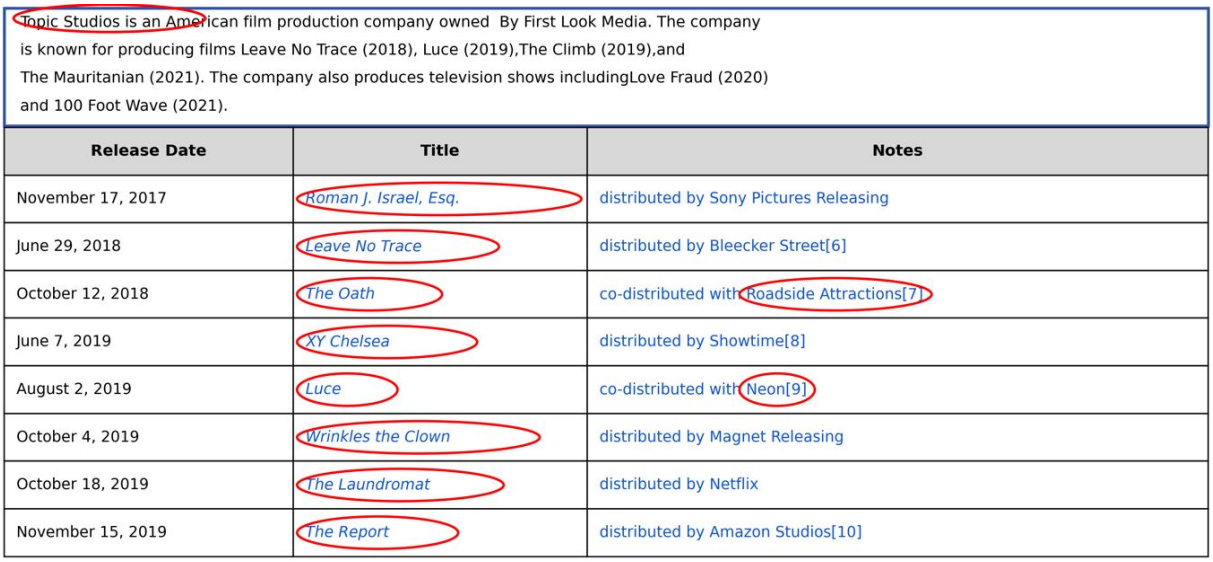
Keywords: Relation extraction, Large Language Models, Information Extraction, Text-Table complementarity

1. Introduction

While documents may contain tables with associated paragraphs, i.e. paragraphs that elaborate on table contents, previous work in Relation Extraction (RE) has typically processed textual content and tables as independent sources, overlooking the semantic relationships that arise from their interaction. Tables in a document rarely convey meaning in isolation; they are generally supported by surrounding text that offers necessary context. Conversely, the narrative text frequently introduces and expands upon information that is synthesized into tables. This strong semantic interplay suggests that considering tables and text jointly rather than separately could significantly improve the quality and completeness of relation extraction. This is illustrated by the example in Figure 1. In the paragraph, we can identify lexically expressed relations such as "`Roman J. Israel, Esq. is a product of the Topic Studio company`", "`Roman J. Israel, Esq. was directed by Dan Gilroy`". The table, in turn, carries relationships expressed between the values in the `Title` and `Notes` columns, such as "`Roman J. Israel, Esq. is distributed by Sony Pictures Releasing`". While the paragraph only mentions one `Topic Studio` production, the table enriches this information by presenting, in chronological order, other productions by the studio (including the one cited in the paragraph). Nevertheless, the name of the production company is not explicitly mentioned in the table. Collaborative relationships between `Topic Studio` and `Roadside Attractions` or `Neon` can also be identified. Only by jointly processing the two modalities (text and tables), one can extract complete and semantically rich relation triples such as: `(LeaveNoTrace, product_of, Topic Studio)`, `(Topic Studio, collaboration, Neon)`.

Ignoring this interplay leads to fragmented or incomplete extractions, ultimately limiting the performance of downstream NLP tasks such as question answering, fact verification, relation extraction, etc. Despite the clear and demonstrated value of extracting relations expressed across different modalities, to the best of our knowledge, only one study has investigated the joint extraction of triples from both text and tables (Zhuang et al., 2022). In that work, a BERT-based model has been trained to identify only one entity, without identifying the type of relations. The lack of research and results in this area can be explained by two main factors. Firstly, each modality (text or table) has its own structural characteristics and therefore requires specific processing methods. Until now, tools have typically

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1341–1351
11–16 May 2026.



Topic Studios is an American film production company owned By First Look Media. The company is known for producing films Leave No Trace (2018), Luce (2019),The Climb (2019),and The Mauritanian (2021). The company also produces television shows includingLove Fraud (2020) and 100 Foot Wave (2021).

Release Date	Title	Notes
November 17, 2017	*Roman J. Israel, Esq.*	distributed by Sony Pictures Releasing
June 29, 2018	*Leave No Trace*	distributed by Bleecker Street[6]
October 12, 2018	*The Oath*	co-distributed with Roadside Attractions[7]
June 7, 2019	*XY Chelsea*	distributed by Showtime[8]
August 2, 2019	*Luce*	co-distributed with Neon[9]
October 4, 2019	*Wrinkles the Clown*	distributed by Magnet Releasing
October 18, 2019	*The Laundromat*	distributed by Netflix
November 15, 2019	*The Report*	distributed by Amazon Studios[10]

Figure 1: Illustration of the need for joint extraction from text and tables. Key triples are only recoverable by combining textual and tabular information (*Source: Wikipedia page "Topic Studios".*)

been trained to perform a single well-defined task. Secondly, no aligned text/table corpus annotated at a fine-grained semantic level is currently available.

In this work, we fill up this gap and we introduce several contributions for improving RE systems:

- The definition of an original relation extraction task in which (i) the subject is an entity and the object can be either an entity or a value, and (ii) the subject and the object must be extracted from two distinct modalities (text and table).
- ReTaT, a manually annotated corpus, drawn from Wikipedia pages in three domains - business, telecommunications and female celebrities - which were selected to ensure diversity in terms of domain, size, relation types, language and size of surrounding text;
- Experiments that confirm the relevance of this task, by computing a measurable performance gain, i.e. by counting the number of original triples identified using jointly tables and text, but not found using tables or text alone;
- The implementation of a baseline system for this task using LLMs' potential, given that LLMs have recently shown great ability to process information from various types of textual structures.

The remainder of this paper is organized as follows. Section 2 proposes a review of related work on relation extraction from text and tables. Sections 3 and 4 respectively describe the resources, the experimental settings, and the results related to our experiments. Section 5 highlights the relevance of the task by computing a performance gain. Finally, Section 6 concludes and outlines future directions.

2. Related work

Many studies described in the literature have focused on extracting semantic relations expressed in textual structures such as sentences (Zhao et al., 2024), paragraphs (Zheng et al., 2023), or tables (Jiang et al., 2022; Yang et al., 2022; Liu et al., 2023), to name a few. Although less numerous, some works do exist that propose approaches capable of reasoning over both text and structured tabular data, such as tables. TaBERT (Yin et al., 2020), an extension of the BERT model (Devlin et al., 2019), is a language model designed to jointly understand natural language and tabular data. It was pre-trained on a corpus of 26 million tables and their textual contexts in English. The performance of TaBERT has been validated on question answering (QA) and SQL query generation tasks, after being fine-tuned using manually annotated QA datasets. Zayats et al. (2021) introduced a novel mechanism to enrich table representations based on the surrounding text, using an adaptation of Graph Neural Networks within a transformer framework. Their model was pre-trained on 1.6 million Wikipedia tables and a corpus of real-world natural language questions. Significant performance gains were observed on QA tasks. ReSel (Zhuang et al., 2022) is a step closer to our setting, as it tackles N-ary relation extraction from scientific documents using both text and tables. The authors pose the N-ary relation extraction task as a question-answering problem: given a query structured as a partial N-ary tuple (e.g., <Task, Method, Dataset, Metric, Score>), a BERT-based model is trained to identify the missing final entity (Score) within the scientific document. Thus, it is indeed a relation extraction

task, but only one entity is retrieved, and the relation itself is not explicitly identified. Although they leverage the semantic relationships between tables and their surrounding text, these approaches were not specifically tailored to the task of extracting relations jointly expressed across the two modalities. Nevertheless, fine-tuning models like TaBERT for this task could yield some initial insights. This highlights the need for a manually annotated corpus of (text, table) pairs, where each relation is expressed jointly in text and tables.

While table-focused datasets (e.g., TabFact (Chen et al., 2019)) emphasize table understanding, entity linking, or question answering, they do not provide manually annotated, gold-standard relation triples grounded in both tables and text. Similarly, hybrid or text-table QA datasets such as HybridQA (Chen et al., 2020b), OTT-QA (Chen et al., 2020a), and FinQA (Chen et al., 2021) target text and table question answering, but are not designed for the explicit extraction of triples with a grounding distributed over both text and tables.

In this work, we introduce ReTaT, a dataset that systematically aligns Wikipedia tables with their semantically related paragraphs, where each (text, table) pair is annotated with the relations jointly expressed in the two modalities.

3. The ReTaT dataset

This section details the construction of ReTaT, a set of three manually annotated datasets containing either full Wikipedia pages with both text and tables, or aligned paragraph-table pairs extracted from Wikipedia pages. The distinct characteristics of these datasets enable a more comprehensive assessment of extraction tools across different configurations. These datasets differ mainly in their domain, language, and size; the set and the type of relations they target; and the size of the text surrounding the table. The different processes involved – selecting domains and relations, collecting Wikipedia pages with tables, annotating triples – are detailed below alongside statistics about these datasets. Some steps were performed in the same way for all three corpora, while others had their own specific methodology. By carefully selecting and annotating linked textual and tabular data from Wikipedia pages, the ReTaT dataset enables controlled experiments for RE from single vs. multiple types of textual structures, setting a novel benchmark for future NLP research.

3.1. Domains and their relation types

Business. The business domain was chosen due to its practical relevance and the abundance of publicly available textual and tabular data, which reflect real-world information extraction challenges. We relied on the CORE dataset (Borchert et al., 2023), a high-quality resource specifically designed for extracting company relations, which are a subset of business relations. CORE is manually annotated, covering the 11 relations listed in Table 1. In this dataset, triples have entities as subject and object.

Female Celebrities. This dataset highlights famous women across various domains. No target list of relations was defined, since we operate in an open relation extraction context. Nevertheless, we restricted the annotation to Wikidata properties. We identified relations by focusing on entities present in the document tables and their mentions in the surrounding text. Because this open approach naturally surfaces a long tail of highly specific or infrequent relations, we retained only relations that occurred at least five times across the dataset to ensure the benchmark evaluates representative semantic patterns. This frequency constraint was exclusively applied to the Female Celebrities domain.

Telecommunications. This dataset centers on documents related to Wi-Fi standards and networking equipment, which are grouped into different themes. No target relations are defined since we operate in an open relation extraction context. Nevertheless, we restricted the annotation to Wikidata properties.

3.2. Collecting Wikipedia pages and constructing Text-Table pairs

Pages were extracted from recent (2024) English (Business and Telecommunications) or French (Female Celebrities) Wikipedia dumps in XML format. Each Wikipedia page was retrieved from the dump as a plain-text WikiText file, where tables follow the native MediaWiki markup syntax (i.e., pipe-delimited WikiText tables) and prose content is encoded as annotated plain text. As part of our preprocessing pipeline, tables were parsed and converted into structured CSV files (one file per table, with rows and columns explicitly delimited), while the surrounding paragraphs were extracted and stored as plain text files. This separation into two distinct file formats enabled systematic downstream alignment between tabular and textual content during the annotation phase.

Business. We collected full Wikipedia articles containing the original CORE sentences to ensure their relevance for the Business domain. From this set, we manually selected 255 articles containing at least one table that was semantically relevant

to the target relations. For each page, we compute cosine similarity using sentence embeddings produced by `all-mpnet-base-v2` (Reimers and Gurevych, 2019), a Sentence-BERT model that has demonstrated strong performance on semantic textual similarity tasks, to select the paragraphs most semantically related to the table content. The most relevant paragraphs are the ones that get higher cosine scores. Subsequently, all automatically generated table–paragraph pairs were manually verified by the annotation team to ensure semantic coherence and contextual relevance. This validation step allowed us to correct mismatches and confirm that both elements referred to the same business-related entities or events. The Business dataset thus comprises 255 high-quality, manually validated table–paragraph pairs.

Female Celebrities. The initial corpus extraction focused on French Wikipedia pages that included both textual and tabular content. From this larger pool, 100 pages were retained based on diversity, informativeness, and structural richness. Preference was given to pages whose tables express relational information with multiple attributes (rather than simple entity lists) and whose surrounding text provides contextual or analytical information relevant to those tables. After selection, each page underwent manual verification to confirm that essential information appeared consistently in both types of textual structures. The dataset was then curated from these 100 French Wikipedia pages containing a total of 665 tables, which can be categorized into four themes: performers and creative artists in arts and entertainment (e.g. Céline Dion; 21 pages), individual athletes and competition lists in sports (e;g. Serena Williams; 36 pages), sovereigns, consorts and dynastic lists within Royalty and Nobility (e.g. liste des duchesses de Lorraine; 33 pages), and biographies of politicians or academics, as well as institutional or thematic lists related to Public Affairs and Academia (e.g. Emmy Noether; 10 pages).

Each candidate page was manually verified to ensure the presence of both rich tabular content and substantive accompanying text suitable for future joint triple extraction. Human verification was performed to confirm that key information appears consistently across both tables and text. This corpus thus offers a valuable resource for research on entity-centric relation extraction and cross-modal information alignment in multilingual contexts.

Telecommunications. This dataset was curated from 32 English Wikipedia pages containing a total of 144 tables. The pages of this corpus can be categorized into four themes: Standards for norms and protocols (e.g. IEEE 802.11ad; 7 pages), Concepts and techniques with technical concepts (e.g. WaveLAN; 6 pages), Equipment for networking devices (e.g. AirPort; 16 pages) and Companies as organizations (e.g. Cisco; 3 pages). Each page was selected to ensure the presence of tabular data, enabling the identification of semantic relationships across text and tables during the annotation phase of the corpus. The corpus was designed to maintain topic consistency, ensuring that all documents revolve around telecommunications, networking, or related domains.

3.3. Annotation process

This annotation aims to identify semantic relations between two items (e1 and e2) that occur across two different textual structures: a textual paragraph and a table or infobox. Each relation is mediated by a predicate linking the two items. e1 is an entity whereas e2 can either be an entity or a litteral attribute. The three components of the relation (e1, predicate, e2) appear within a shared context, although they are not necessarily located within the same paragraph or table — they must be distributed across both structures. Annotation was supported by a custom web-based interface that enabled synchronized highlighting of text spans and table cells. Where applicable, entity mentions were canonicalized by linking them to Wikidata QIDs to maintain consistency across texts and tables and support downstream processing. We used the following guidelines during the annotation process: (i) Entity mentions were linked to Wikidata QIDs to ensure consistency across text and tables. (ii) Annotations were performed to capture relationships distributed across text and tables that are close in a document. (iii) Predicates were selected from the curated list of Wikidata properties to ensure semantic coherence.

As a result, the corpus is distributed as paired files sharing a unique document identifier: a plain text file for the extracted paragraphs and a CSV file for the structured table. Gold-standard relation triples are provided separately in a dedicated annotation file, indexed by document identifier to allow direct alignment with the source content.

Business. Four annotators conducted the annotation: one graduate student with a dual background in computational linguistics and business, and three experts in information extraction. They underwent a training phase on a set of 50 documents (paragraphs and tables), including adjudication steps. These sessions allowed the team to collaboratively refine the annotation guidelines and decision rules. Then each document was annotated by two annotators and systematically verified by an expert, resulting in 2,997 triples. Considering that two triples match when the textual spans of their entity mentions overlap exactly and if they

share the same relation type and direction, the inter-annotator agreement on the double-annotated corpus is $k = 0.70$ (Cohen's kappa).

Female Celebrities. Two annotators performed the annotations: a PhD student specializing in knowledge engineering and a research engineer in information extraction. A double-annotation protocol was applied to 100% of the dataset: the PhD student conducted the primary annotation, and the research engineer independently reviewed and re-annotated the same documents. This procedure ensured full cross-validation and expert verification of the entire corpus.

Telecommunications. A single annotator, a senior researcher in AI with ten years of experience in a Telco company, performed the annotations. The annotator conducted a first round of annotation on the entire corpus, followed by a thorough review of all annotations a few weeks later to ensure accuracy and coherence across the dataset. It is planned to engage a second reviewer in the future to enhance the robustness of the annotation process.

3.4. Descriptive statistics

Tables 1, 2 and 3 give the list of relations and their distribution in each of the annotated datasets.

The Business dataset exhibits a coherent relational structure, mainly centered on corporate relations such as *product_or_service_of* (1,669 instances), *client_of* (526), and *collaboration* (384). All relations are object properties linking organizations or corporate entities, with no data properties (i.e., literal attributes). This clear relational scope explains the strong extraction results obtained in this dataset.

In contrast, the Female Celebrities dataset contains a wider and more heterogeneous range of social and professional links, including *noble title* (526), *award received* (425), and *participant in* (314). The distribution is more balanced, with roughly 65% object properties (e.g., *spouse, creator*) and 35% data properties (e.g., *ranking, vote received*).

The Telecommunications dataset is the most diverse and numerically driven, with relations like *data transfer speed* (559), *uses* (127), *order number* (64), and *frequency* (61) reflecting technical attributes. Here, data properties dominate (70% of relations), while object properties such as *parent organization or unit* or *developer* make up the remaining 30%.

Overall, the datasets differ not only in size but also in the balance between object and data properties, providing complementary conditions for evaluating joint text–table relation extraction and testing model adaptability across domains of increasing structural and semantic diversity.

4. Experimental settings and results

We conducted experiments using Large Language Models (LLMs), at a time when LLMs have demonstrated their ability to extract relations from either text or tables in isolation. Our goal was not to design an efficient model but rather to assess the quality and suitability of the ReTaT datasets for this new joint extraction task.

4.1. Selected LLMs

We conducted experiments using two recent LLMs: **llama-4-maverick-17b-128e-instruct**, a 17-billion-parameter transformer optimized for multi-turn instruction following and factual reasoning, and **deepseek-r1-distill-llama-70b**, a distilled variant from a larger 70-billion-parameter model, focusing on extracting reusable knowledge representations through high-fidelity distillation from expert LLMs. For all experiments, we used 4-bit precision on RTX 8000 GPUs, ensuring reproducibility via fixed random seeds and standardized decoding hyperparameters (temperature = 0.2, max_tokens = 2048).

4.2. Instruction-based Prompt for Joint Relation Extraction

We used an instruction-based prompting approach with a prompt explicitly tailored for joint relation extraction, ensuring that the model captures the interconnections between narrative text and structured table content. The prompt begins by precisely framing the objective (joint extraction), from both a text and a table content. To maximize consistency and downstream usability, the expected output follows a strictly defined triplet schema. Each model was evaluated under zero-shot and few-shot configurations. For the few-shot setup, the prompt includes 3 examples representing distinct relation types. Below is the finalized instruction template employed across both zero-shot and few-shot experiments:
The prompt presented below is the result of a systematic exploration of several prompt formulations. We experimented with variants that progressively incorporated additional constraints, notably the cross-source grounding rule (requiring each triple to draw evidence from both text and table) and the internal validation instruction. The version retained here consistently yielded the best results across all three datasets and both evaluation configurations, and is therefore used in all reported experiments.
PROMPT:

Property Label	Description	# #
Product or service of	e_1 is offered for commercial distribution by e_2.	1669
Client of	e_1 uses (and pays for) products or services from e_2.	526
Collaboration	e_1 and e_2 collaborate in parts of their business activities.	384
Subsidiary of	e_2 legally owns e_1.	174
Acquired by	e_2 purchases a controlling stake in e_1.	81
Shareholder of	e_1 owns shares in e_2.	61
Brand of	e_2 offers products or services of e_1 (brand).	40
Traded on	Shares of e_1 are listed on e_2 (stock exchange).	5
Merged with	e_1 and e_2 merged their business operations (fully or partially).	1

Table 1: Semantic relation types in the Business dataset with the CORE definitions and distribution.

PID	Property Label	# #
P97	noble title	526
P166	award received	425
P1344	participant in	314
P1352	ranking	140
P26	spouse	108
P1111	votes received	104
P1346	winner	92
P170	creator	81
P8810	parent	24
P184	doctoral advisor	16
P39	position held	6

Table 2: Semantic relation types and number of occurrences in the Female Celebrities dataset.

You are an expert in relation extraction. For this task, think step by step for each given relation.
*Return **only** a valid JSON array of arrays, where each inner array is exactly: ["subject", "relation", "object"]*
Rules:
- The provided text is contained within the <TEXT></TEXT> tags. The provided tables are contained within the <TABLE></TABLE> tags.
- Each of the subject, predicate, and object must come from the provided text and/or tables.
- Keep a triple only if it uses information from both the text and the table. Reject any triple that is produced using only the text or only the table.
- "relation" MUST be one of these identifiers: relations_list. Do not invent new relation labels.
- Output MUST be valid JSON. No prose, no code fences, no prefixes/suffixes, no trailing commas.
- Do not include any explanation or text outside the JSON.
- Before outputting, perform an internal validation pass: for every triple you plan to output, identify the source of each element (TEXT or TABLE). If any triple fails rule (3) or (4), remove it. After validation, if no triple remains, output '[]'.
List of relations with their description: relations_text
Now process the new data: {data}

PID	Property Label	# #
P6711	data transfer speed	559
P2283	uses	127
P8470	order number	64
P2144	frequency	61
P9767	edition/version	58
P2928	storage capacity	52
P1343	described by source	49
P880	CPU	48
P13351	model number	45
P13525	RAM capacity	42
P5204	commercialization date	40
P2669	discontinuation date	32
P306	operating system	31
P2284	price	30
P31	instance of	28
P2664	units sold	27
P2403	total assets	23
P2139	total revenue	23
P2560	GPU	22
P2295	net profit	22
P2149	clock frequency	19
P2049	width	19
P1128	employees	18
P2109	nominal power output	18
P2652	partnership with	15
P2048	height	14
P2067	mass	13
P749	parent organization or unit	13
P12323	working memory type	13
P1056	product or material produced	12
P5524	horizontal depth	11
P169	chief executive officer	8
P516	powered by	7
P2043	length	6
P178	developer	5

Table 3: Semantic relation types and number of occurrences in the Telecommunication dataset.

4.3. Evaluation metrics

To comprehensively assess the performance of the LLMs on the joint text-table relation extraction task, we defined three evaluation levels that progres-

sively relax the matching criteria. This approach provides a more nuanced understanding of how models behave, that goes beyond strict triplet exactness, by highlighting areas of partial understanding and relational consistency.

- **Exact Triple Match:** A triple is considered *correct* only if all three components (subject, predicate, object) exactly match a reference triple in the document gold standard.
- **Entity-Level Partial Match:** Here, a triple is considered *partially correct* if the predicted subject and object match a gold-standard pair, regardless of the relation label. This relaxed criterion isolates the model's ability to correctly identify relevant subject-object pairs across text and table.
- **Relation-Level Partial Match:** A prediction is considered *partially correct* if the extracted relation and either the subject or the object match a reference triple in the gold document. This dimension measures whether the model infers the correct relation type even when identifying the entity boundaries is imperfect.

Each evaluation level relies on standard information extraction metrics: Precision, Recall, and F1-score.

4.4. Results and discussion

Table 4 presents the results obtained by LLaMA-4 and DeepSeek models across the three manually annotated datasets. Results are reported for both zero-shot and few-shot (3-shot) configurations under three evaluation granularities: (i) exact triplet match, (ii) partial match on both entities, and (iii) partial match on relation plus one entity.

Across all corpora, the few-shot setting substantially improves precision, recall, and F1-score, confirming that limited in-context examples help models internalize structural constraints and relational semantics. DeepSeek outperforms LLaMA-4 in both precision and recall, denoting greater stability and control in structured reasoning.

The *Business dataset* yields the highest scores overall, with DeepSeek reaching 0.43 F1 (exact) and 0.47 F1 (2-entity partial) in the few-shot setup, while LLaMA-4 attains 0.31 and 0.34 F1, respectively. This good performance is largely due to the dataset's well-defined relation schema, explicit text–table alignments, and lexically stable entities such as company names and events. The clear semantic overlap between narratives and structured information enables both models to efficiently leverage complementary textual cues. DeepSeek's higher precision (0.71) reflects its ability to make conservative and accurate predictions, whereas LLaMA-4 shows slightly broader but noisier recall.

By contrast, the *Female Celebrities dataset* exhibits the lowest performance levels, with F1 rarely exceeding 0.15 for LLaMA-4 and 0.42 for DeepSeek under partial-matching criteria. Several factors explain this gap: the corpus is smaller, with fewer relational examples; it lacks explicit text–table pairings, increasing contextual noise; the texts are much longer, diluting relevant signals; and the relation list is considerably larger and more diverse, making fine-grained distinctions harder. Despite these difficulties, DeepSeek's few-shot configuration shows a notable improvement (up to 0.42 F1 under relation + 1 entity partial matching), indicating that large models can still recover plausible relational patterns in noisy environments.

The *Telecommunications dataset* produces intermediate results between the two extremes. In few-shot mode, DeepSeek achieves 0.38 F1 (relation + entity partial) and LLaMA-4 reaches 0.26 F1 under the same measure. These moderate outcomes result from the corpus characteristics: smaller size limiting relational coverage, tables dominated by numeric entries that provide weak lexical anchors, the relations involves highly technical and fine-grained ambiguities, especially compared to other domains. For instance, the model must distinguish between closely related software versions and hardware models which share overlapping features, similar naming conventions, and incremental updates. Such distinctions require domain-specific knowledge and precise contextual understanding, making classification more challenging than in domains with broader or more distinct categories.

Overall, the results also highlight the importance of developing dedicated datasets like ReTaT to study joint text–table extraction, emphasizing that combining structured and unstructured textual information is crucial for achieving broader relational coverage and more holistic knowledge extraction.

4.4.1. Cross-Corpus Comparison

Aggregated across domains, several conclusions emerge. First, few-shot prompting consistently enhances performance, demonstrating the adaptability of in-context learning for structured reasoning. Second, dataset characteristics strongly influence accuracy. The Business corpus, with its clear schema and alignment, produces the most reliable results, whereas the smaller and noisier Female Celebrities and Telecommunications corpora yield much lower precision and recall. The larger relation inventories and the presence of numeric or overly long text segments introduce further noise, explaining the significant precision gap (Business $>$ 0.47 F1 vs. $<$ 0.30 for the other datasets).

4.4.2. Error Analysis

Error inspection reveals recurring challenges: **Entity boundary mismatches**, overly extended or truncated entity spans cause misalignment with gold annotations; **Relation confusion**, models alternate between semantically neighboring predicates; and **Single source extraction bias,** where the model infers relations solely from text or table, rather than jointly from both.

5. Is Joint Extraction Necessary?

To prove the usefulness of the task and the ReTaT dataset, we measure the contribution of a joint extraction approach, compared to a text-only or table-only approach. This contribution represents the percentage of additional correct triples identified by the joint extraction method compared to the single-source settings (text-only and table-only), computed as follows:

$$C = \left(\frac{\text{\#NewTriples}}{\text{\#Triples Separately } + \text{\#NewTriples}} \right)$$

where,
- `NewTriples` are correct triples identified uniquely through joint reasoning across text and tables (i.e., not present in triples extracted separately from text or tables).
- `TriplesSeparately` is the sum of correct triples retrieved independently from text and tables.

For this purpose, we systematically derived three corpora from the Business dataset, which is the larger one. All three corpora are based on the same set of 255 text and tables, with different annotations:

- **CorpusP (Text-only):** Only the 255 selected paragraphs were manually annotated with relations explicitly expressed in text only, leading to a total of 239 triples.
- **CorpusT (Table-only):** Relations explicitly expressed in tables only were manually annotated in the 255 selected tables, by inspecting the interactions of the rows and columns. This resulted in 855 triples.
- **CorpusPT (text-table jointly):** This is the Business dataset presented in Section 3 composed of 2,997 triples. It ensures that each relation is supported jointly by evidence from both text and tables.

Thanks to a unified annotation schema across the various types of textual structures, the corpora share the same entity space (business organizations, products, and actors) and relations, which guarantees the possibility to compare results from the three of these corpora. To quantify the benefit of using joint information from both text and tables, we measured the contribution of our approach compared to using each textual structure in isolation. In the annotated business dataset, there are 1,034 triples in either texts or tables in isolation, whereas 2,544 triples were discovered from both a text and a table (C=70,7%), demonstrating that more than two-thirds of the correct triplets could *only* be captured when models jointly considered both textual structures.

To evaluate whether large language models (LLMs) also benefit from joint structured reasoning, we applied the same analysis using the best-performing model from previous experiments, DeepSeek-70B. When restricted to text-only or table-only settings, the model correctly extracted 655 triplets under exact matching. However, when allowed to jointly process both text and table content, it identified 788 correct triplets, yielding a contribution of 54.6%.

6. Conclusion and Future Work

This paper introduced ReTaT, a novel benchmark designed to foster research on joint relation extraction from text and tables. Unlike previous datasets that treat these two sources independently, ReTaT explicitly aligns textual and tabular evidence, allowing the study of relational reasoning that emerges only through their interplay. Moreover, relations in ReTat are jointly expressed across text and tables, which is complementary to the relations available in each of these modalities.

We manually curated and annotated three domain-specific datasets; Business, Telecommunications, and Female Celebrities, ensuring diversity in domains, sizes as well as relation numbers and types. Experiments with large language models (LLaMA-4 and DeepSeek-R1-70B) demonstrated that combining structured and unstructured textual evidence leads to substantially more complete and semantically rich extractions, with more than 70% of valid triples being discoverable only when both modalities are jointly considered.

Beyond its immediate use as a benchmark, ReTaT paves the way for several directions of future work. First, while this study focused on LLM-based approaches, an important next step will be to evaluate non-LLM models specifically designed for joint text–table reasoning, such as TaBERT (Yin et al., 2020), which was pre-trained on large collections of Wikipedia tables paired with their textual context. Comparing LLM-based and fine-tuned encoder-based approaches on ReTaT will provide a more comprehensive picture of the state-of-the-art on this task. Second, future work could explore the use of similarity-based text–table alignment strategies, as used in the Business corpus, to improve dataset construction and reduce annotation noise across

Business Dataset							
...		Zero-shot			Few-shot (3)		
Model	Eval. Method	P	R	F1	P	R	F1
LLaMA-4	Exact matching	0.23	0.11	0.15	0.48	0.22	0.31
	Partial matching (2 entities)	0.29	0.14	0.18	0.53	0.25	0.34
	Partial matching (relation + 1 entity)	0.44	0.21	0.28	0.69	0.32	0.44
DeepSeek-70B	Exact matching	0.61	0.26	0.37	0.71	0.31	0.43
	Partial matching (2 entities)	0.64	0.28	0.39	0.79	0.34	0.47
	Partial matching (relation + 1 entity)	0.68	0.30	0.41	0.76	0.33	0.46
Female Celebrities Dataset							
...		Zero-shot			Few-shot (3)		
Model	Eval. Method	P	R	F1	P	R	F1
LLaMA-4	Exact matching	0.07	0.05	0.06	0.06	0.04	0.05
	Partial matching (2 entities)	0.15	0.11	0.12	0.13	0.09	0.11
	Partial matching (relation + 1 entity)	0.19	0.14	0.16	0.13	0.09	0.10
DeepSeek-70B	Exact matching	0.12	0.10	0.11	0.16	0.18	0.17
	Partial matching (2 entities)	0.19	0.16	0.18	0.19	0.20	0.19
	Partial matching (relation + 1 entity)	0.31	0.26	0.29	0.42	0.42	0.42
Telecommunications Datasets							
...		Zero-shot			Few-shot (3)		
Model	Eval. Method	P	R	F1	P	R	F1
LLaMA-4	Exact matching	0.20	0.07	0.11	0.38	0.13	0.20
	Partial matching (2 entities)	0.32	0.12	0.18	0.50	0.18	0.26
	Partial matching (relation + 1 entity)	0.31	0.12	0.17	0.51	0.18	0.26
DeepSeek-70B	Exact matching	0.21	0.08	0.11	0.39	0.27	0.32
	Partial matching (2 entities)	0.28	0.10	0.15	0.43	0.29	0.35
	Partial matching (relation + 1 entity)	0.29	0.11	0.16	0.46	0.32	0.38

Table 4: Zero-shot and few-shot results for LLaMA-4 and DeepSeek models.

domains. Finally, ReTaT opens broader research avenues in cross-structured information extraction, knowledge base population, and explainable reasoning, where jointly exploiting structured and unstructured textual evidence is key to achieving more complete and semantically grounded knowledge extraction.

Acknowledgments

This work was supported by the French National Research Agency (Agence Nationale de la Recherche - ANR) under the ECLADATTA project (ExtraCtion of LAtent knowledge in Documents by conjointly Analyzing Texts and TAbles, https://ecladatta.github.io/), grant number ANR-22-CE23-0020 (2023–2026). We also would like to thank Anaïs Schlosser who participed in the annotation process.

7. Bibliographical References

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2020. TURL: table understanding through representation learning. *CoRR*, abs/2006.14806.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Muhammad Irfan and Liam Murray. 2023. Micro-credential: A guide to prompt writing and engineering in higher education: A tool for artificial

intelligence in llm. Technical report, University of Limerick.

Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. OmniTab: Pretraining with Natural and Synthetic Data for Few-shot Table-based Question Answering. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL*, pages 932–942. Association for Computational Linguistics.

Xingzuo Li, Kehai Chen, Yunfei Long, and Min Zhang. 2024. LLM with relation classifier for document-level relation extraction. *CoRR*, abs/2408.13889.

Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. 2023. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *J. Web Semant.*, 76:100761.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Arif Shahriar, Rohan Saha, and Denilson Barbosa. 2023. Relational extraction on wikipedia tables using convolutional and memory networks.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. In *60th Annual Meeting of the Association for Computational Linguistics ACL*, pages 528–537. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *58th Annual Meeting of the Association for Computational Linguistics ACL*, pages 8413–8426. Association for Computational Linguistics.

Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. 2021. Representations for question answering from documents with tables and text. In *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL*, pages 2895–2906. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

Lei Zhao, Ling Kang, and Quan Guo. 2025. Zero-shot document-level biomedical relation extraction via scenario-based prompt design in two-stage with llm. *arXiv preprint arXiv:2505.01077*.

Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. 2024. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Comput. Surv.*, 56(11).

Yifan Zheng, Yikai Guo, Zhizhao Luo, Zengwen Yu, Kunlong Wang, Hong Zhang, and Hua Zhao. 2023. A survey on document-level relation extraction: Methods and applications. In *Proceedings of the 3rd International Conference on Internet, Education and Information Technology (IEIT 2023)*, pages 1061–1071. Atlantis Press.

Yuchen Zhuang, Yinghao Li, Junyang Zhang, Yue Yu, Yingjun Mou, Xiang Chen, Le Song, and Chao Zhang. 2022. Resel: N-ary relation extraction from scientific text and tables by learning to retrieve and select. In *Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 730–744. Association for Computational Linguistics.

8. Language Resource References

Philipp Borchert and Jochen De Weerdt and Kristof Coussement and Arno De Caigny and Marie-Francine Moens. 2023. *CORE: A Few-Shot Company Relation Classification Dataset for Robust Domain Adaptation*. Association for Computational Linguistics.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2020a. Open question answering over tables and text.

Wenhu Chen and Hongmin Wang and Jianshu Chen and Yunkai Zhang and Hong Wang and Shiyang Li and Xiyou Zhou and William Yang Wang. 2019. *TabFact: A Large-scale Dataset for Table-based Fact Verification*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics (EMNLP)*, pages 1026–1036. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Kenneth Huang, Bryan R. Routledge, and William Yang Wang. 2021. Finqa: A dataset of numerical reasoning over financial data.